TECHNISCHE
UNIVERSITÄT
DRESDEN

Application of the Profile-Likelihood Method to the Combination of Cross Section Measurements

Bachelor-Arbeit
zur Erlangung des Hochschulgrades
Bachelor of Science
im Bachelor-Studiengang Physik

vorgelegt von

Konstantin Schubert
geboren am 19.12.1990 in Saarbrücken

Institut für Kern- und Teilchenphysik
Fachrichtung Physik
Fakultät Mathematik und Naturwissenschaften
Technische Universität Dresden
2012

Eingereicht am 21. Mai 2012

1. Gutachter: Prof. Dr. Michael Kobel
2. Gutachter: Prof. Dr. Kai Zuber

Summary

Abstract
English:

We use the method of maximum likelihood to obtain a combined value for the $Z \to \tau\tau$ cross section from measurements in three different final states. The profile likelihood method is applied to find confidence intervals on the stated result. We carefully justify the form of the employed likelihood function and discuss the frequentist foundations of the applied methods. The necessary algorithms are implemented with the RooFit[VK03] framework and tested on simple examples. The Neyman construction and the unified approach are demonstrated in order to motivate the more sophisticated test statistic used in the combination.

Abstract
Deutsch

Wir nutzen die Maximum-Likelihood-Methode um Messungen des $Z \to \tau\tau$ Wirkungsquerschnitts aus drei verschiedenen Endzuständen zu einem gemeinsamen Wert zu kombinieren. Mithilfe der Profile-Likelihood-Methode ermitteln wir für das Endergebnis ein Konfidenzintervall. Wir begründen die funktionale Form der gewählten Likelihood-Funktion und besprechen ausführlich die frequentistischen Grundlagen der angewendeten Methoden. Der verwendete Algorithmus wurden unter Verwendung des RooFit[VK03] Frameworks programmiert. Wir demonstieren ihn anhand einfacher Beispiele. Als Motivation für die Wahl der später verwendeten Teststatistik werden die die Neyman-Konstruktion und der Unified-Approach besprochen.

Contents

List of Figures

List of Tables

1 Introduction

Statistical methods have always played an important role in experimental particle physics. While the research aims at examining smaller and smaller effects, the luminosity of modern particle colliders is growing in large steps. There is no doubt that more and more sophisticated techniques will be necessary to answer future physical questions. Since the birth of modern physics, Bayesians and frequentists are debating the best procedure for the evaluation of measured data. The aim of this work is to demonstrate the application of important frequentist methods with a special focus on the Method of Maximum Likelihood and related test statistics. Following this introduction, the second chapter will deal with basic concepts of frequentist inference in particle physics. It provides the theoretical foundation and motivation for the methods that are applied later in this work, with a special focus on the profile likelihood ratio. We will carefully demonstrate how the Neyman-Pearson lemma, which originally considers only the comparison of two concurring hypotheses, can indirectly be used to justify the maximum likelihood ratio test statistic as an optimal tool for the setting of limits. Considering nuisance parameters, the latter will be generalized to the profile likelihood ratio. The third chapter is intended to illustrate these abstract concepts with important examples. The limitations of the Neyman construction and of the unified approach will be examined and techniques that are later used for the combination of cross sections will be motivated and tested exemplary. In the fourth chapter, we use the Method of Maximum Likelihood to combine measurements of the $Z \to \tau\tau$ cross section from the three final states $\tau_e\tau_h$, $\tau_\mu\tau_h$ and $\tau_\mu\tau_e$. The combination bases on the results of an ATLAS[AAA$^+$08] conference note[Con12]. After sketching the process of event selection, the systematic uncertainties will be summarized. Using only frequentist arguments, we find a form for the likelihood function that can be applied to the measured cross sections and their systematic uncertainties as they are stated in [Con12]. For the combined cross section, the value $\sigma = 0.88\ nb$ is obtained. The one-sigma confidence intervals are evaluated using the profile likelihood ratio test statistic and calculated to: $[0.80\ nb, 0.96\ nb]$ $(total)$, $[0.87\ nb, 0.89\ nb]$ $(stat)$. We will discuss systematic influences on this result and propose ways to further improve the combination.

2 Theoretical Overview

2.1 The Bayesian and the Frequentist Way to Understand Probability

There are two fundamentally different ways to interpret the mathematically well-defined[Cow98, chap. 1.1] concept of a probability: The *classical*, *frequentist* way and the *Bayesian* (or *subjectivist*) interpretation. In the frequentist interpretation, a probability γ makes a statement about *relative frequencies of the outcome of an experiment in the large sample limit.* Assume an experiment, that, if repeated infinitely often, gives the result A in a fraction γ of times. If we know this, we are able to state that within each execution of the experiment, we will measure A with a classical probability of γ. On the other hand, the bayesian interpretation defines probability as a personal *degree of belief*, which is stated on a scale from 0 to 1. It is a correct bayesian statement to say: "I believe with 80% that there is a god". This bayesian interpretation is subjective by its definition. It therefore offers the physicist the freedom to incorporate personal judgements about the plausibility of physical models or hypotheses. The lack of objectivity is often criticized as an unacceptable flaw of the Bayesian approach. Many physicists believe that one should only make frequentist statements when publishing results. They call themselves "frequentists" in constrast to the "Bayesians" who prefer the Bayesian interpretation and state a probability distribution for the unknown but fixed parameter of interest. Some efforts have been made to counter the frequentist criticism by making objective bayesian statements. In particular, statisticans tried to find a "non-informative" prior PDF [FC98] for the application of Baye's theorem. But *"In our view, the attempt to find a non-informative prior within Bayesian inference is misguided. The real power of Bayesian inference lies in its ability to incorporate 'informative' prior information, not 'ignorance'."*[FC98]

It can hardly be stressed enough that both the Bayesian and the frequentist interpretation are correct. They are both logically consistent with the laws that mathematically define probability. Also, both are useful in the scientific context, and they should be chosen depending on the intention of the statement. As Cowan [Cow07] states it: *"Scientists should not be required to label themselves as frequentists or Bayesians. The two approaches answer different but related questions, and a presentation of an experimental result should naturally involve both. Most of the time, one wants to summarize the results of a measurement without explicit reference to prior probabilities; in those cases the frequentist approach will be most visible."*

In the following, we will concentrate on frequentist methods of inference.

2.2 The Experiment

The goal of most physical experiments is to make statements about the value of a fixed but unknown parameter μ. Therefore, the physicist conducts an experiment which leads to a measurable outcome n. In an ideal world, different values of μ would lead to different outcomes n such that by measuring n, $\mu(n)$ could be derived directly.

Unfortunately, due to the statistical nature of the underlying physical processes, this is not the case for most experiments in particle physics. Here, the measured quantity n of the experiment is statistically distributed. Its distribution is described by a *Probability Density Function(PDF)* $f(n|\mu)$, which depends on μ as a parameter and is otherwise defined in the space of possible experimental outcomes, called the *sample space*. Obviously, we can't directly derive the true value μ from the outcome of the experiment, but we can try to make statistical statements that somehow qualify μ.

Despite the fact that n and μ are denoted as scalars, they might equally well represent a vector of different quantities. In practice, scientists often face an armada of unknown parameters. To gather more infomation about μ, most experiments are conducted $N > 1$ times, leading to a list of *independent and identically distributed* (*iid*) random variables $[n_1, ..n_N]$. A realization of this list is called a sample, and N is the sample size. The probability density to measure the values in the sample is then given by $\prod_{i=1}^{N} f(n_i|\mu)$.
Note that we could equivalently define a new N-dimensional sample space with the vector $\vec{n} = (n_1, ..., n_N)$ as the outcome of a single experiment described by $g(\vec{n}|\mu) = \prod_{i=1}^{N} f(n_i|\mu)$. Hence, when "$n$" is written in the following, it can always be understood also as a vector of different outcomes of the same experiment.

2.3 An important Tool: Test Statistics

A function $h(n)$ that is defined on the sample space of the experiment is called a *statistic*. Its value h is a random variable and the probability distribution $g(h(n)|\mu)$ is determined by the function $h(n)$ and the PDF of n. *Test statistics* are statistics that are designed as indicators for different values of the parameter of interest μ. By suitably defining them, one achieves that their PDFs have minimal overlap for different values of μ. Consider an example where we want to distinguish between two possible values μ_0 and μ_1. Be $t(n)$ our test statistic. If we know $f(n|\mu)$, we can find the probability distribution g of t: $g(t|\mu)$. This situation is illustrated in figure (2.1) for the one-dimensional case. We decide to reject the hypothesis μ_0 whenever the test statistic exceeds a certain threshold t_{cut}. The region of values lower than t_{cut} is referred to as the *acceptance region* A, whereas its complement is known as the *critical region* C[1]. With this approach, the probability α to *reject* μ_0, *if it is true,* is

[1]It is not essential to define the acceptance region as the interval of all values $t < t_{cut}$. One can rather choose any subset of the parameter space. In sections 2.7, we will discuss methods of making the optimal choice.

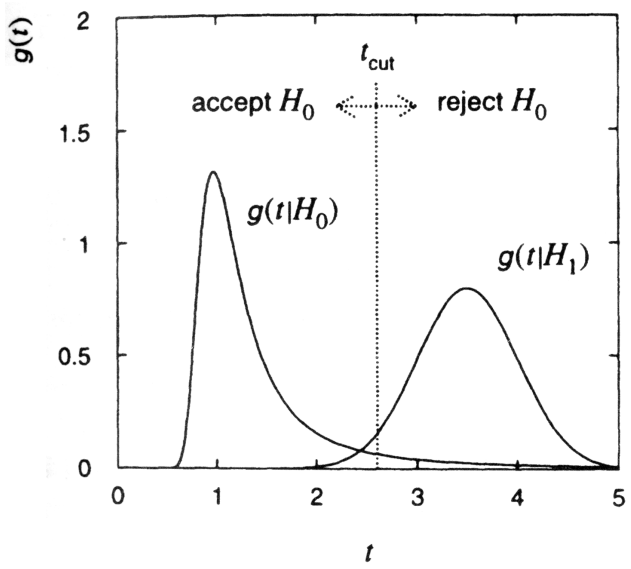


Figure 2.1: The probability distributions of $t(n)$ for the two concurring hypotheses. H_0 represents μ_0, H_1 represents μ_1. Figure taken from [Cow98]

given by:

$$\alpha = \int_C g(t|\mu_0) \ . \tag{2.1}$$

α is the *significane level* of the test, $1-\alpha$ the *efficiency*. On the other hand, the probability β to accept μ_0 if in fact μ_1 is realized, equals:

$$\beta = \int_A g(t|\mu_1) \ . \tag{2.2}$$

The quantity $1-\beta$, which we want to be as big as possible, is then referred to as the *power* of the test.

2.4 Estimators

Another very important tool to narrow down an unknown parameter are *estimators*. An estimator $\hat{\mu}(\vec{n})$ [2] for μ is a statistic that is designed to *approximate* the true value of μ *in the large sample limit*. This approximation is to be understood as a convergence in the sense of probability:

$$\forall \varepsilon > 0 \lim_{N\to\infty} P(|\hat{\mu} - \mu| > \varepsilon) = 0 \ . \tag{2.3}$$

Here, N is the sample size and P denotes the probability that the expression in brackets is true. An estimator that meets this requirement is called *consistent*. Additionally, one often wishes that the

[2] The notation $\vec{n}$ is chosen instead of n to emphasize that the estimator usually depends on a sample of iid outcomes of the same experiment. For counting experiments, this might be counterintuitive since the estimators for mean and variance only depend on the cumulative count.

estimator is also *unbiased*. For an unbiased estimator, the expected value equals the true but unknown parameter:

$$E(\hat{\mu}) = \mu \ . \tag{2.4}$$

If this is not the case, the difference b

$$b = \mu - E(\hat{\mu}) \tag{2.5}$$

is called the bias of the estimator.

2.5 Finding Estimators with the Method of Maximum Likelihood

The Maximum Likelihood Method provides an easy approach to find an estimator for a parameter of interest. We therefore define the *likelihood function* L_n. It states the likelihood to obtain a fixed measurement n depending on the parameters of the experiment and can be obtained simply by inserting n into the PDF $f(n|\mu, \theta)$ of the experiment.

$$L_n(\mu, \theta) = f(n|\mu, \theta) \quad . \tag{2.6}$$

When maximizing L_n, we obtain the values $\hat{\mu}$, $\hat{\theta}$ with the highest probability to create the outcome n. It turns out that for weak conditions, $\hat{\mu}$ and $\hat{\theta}$ are estimators for the true values μ, θ. These estimators are usually unbiased in the large sample limit. The method of maximum likelihood also provides a way to find the variance of the estimator and comes with many more beneficial properties. We will not discuss them any further, as well as we do not give proofs for the statements above. The interested reader is referred to Cowan[Cow98], who gives a good introduction into differenct aspects. In chapter 4, the maximum likelihood method will be applied to calculate the best estimator for the combined cross section.

2.6 Confidence Intervals

Since μ is unknown but fixed, we cannot make frequentist statements about μ itself. We should not try to find the probability for μ to be contained in the interval $[\mu_1, \mu_2]$. It is either always in or never[3]. What we can do is to define a *confidence interval*. This is an interval[4] $\mathcal{I}(n) := [\mu_1(n), \mu_2(n)]$ which is constructed depending on the outcome n of the experiment, in such way that it will contain the true value μ in a fraction γ of infinitely often repeated executions of the experiment. In other words: We

[3] One might argue that with the help of a fictional infinite set of universes with independent physical laws, there does actually exist a freqentist interpretation of the "Bayesian posterior PDF". However, since subjectivity is not allowed, it seems impossible to find a frequentist "bayesian prior".

[4] Note that even though it is called a "confidence interval" it does not necessarily have to be an interval. The arguments given above and in the following can be generalized to I being a subset of the parameter space. This is especially important for the multi-dimensional case.

find a statistic $\mathcal{I}(n)$, with the property that $P(\mu \in \mathcal{I}(n)) = \gamma$. γ is then called the *coverage probability* or the *confidence level*.

2.6.1 General construction of confidence intervals

The goal is to construct a confidence interval from the outcome of the experiment such that it covers the unknown but fixed true value of μ, which will be denoted μ' in the following, in a fraction γ of all experiments. In the most general case, we therefore choose a test statistic for every μ and find its probability distribution: $f(t_\mu|\mu')$. For every μ, we choose an acceptance region A_μ out of the codomain of t_μ, such that if μ was the true value, the probability to find t_μ in A_μ would be γ:

$$\int_{A_\mu} dt_\mu f(t_\mu|\mu) = \gamma \ . \tag{2.7}$$

The confidence interval $\mathcal{I}(n)$ is then defined as the set of all μ, for which $t_\mu(n)$ is in A_μ.

$$\mathcal{I}(n) = \{\mu : t_\mu \in A_\mu\} \ . \tag{2.8}$$

This so-defined set then meets the definition above: $t_{\mu'}(n)$ will be found in $A_{\mu'}$ with probability γ, thus μ' will be contained in $\mathcal{I}(n)$ with probability γ.
The *Neyman construction* in the next chapter (3.2) is a basic example for such an interval construction.

2.7 The Neyman-Pearson Lemma

In section 2.3, we discussed how test statistics can be used to compare two competing hypotheses. Given a certain efficiency, one might wonder how one should construct the acceptance set A in an optimal way. The answer is given by the *Neyman-Pearson-Lemma*. It states that for a given significance α and a given test statistic t, the power of the test is maximized if the acceptance region A is composed by the elements of t with the biggest Likelihood Ratio $\frac{g(t|\mu_0)}{g(t|\mu_1)}$. In other words:

$$A = \left\{ t : \frac{L_{t(n)}(\mu_0)}{L_{t(n)}(\mu_1)} > c \right\} \ , \tag{2.9}$$

where c depends on the desired significance.

Equivalent to (2.9), one can define a new test statistic $r = \frac{g(t|\mu_0)}{g(t|\mu_1)}$ and apply a cut at $r = c$.

2.8 The Likelihood Ratio

Note that the Neyman-Pearson lemma does not give any advice how the test statistic t should be chosen initially. It is obvious that a bad choice of t can make the distinction between the two

hypotheses impossible if it eliminates the necessary information. A test statistic that saves all the information about the experimental outcome is the measurement n itself. Therefore it seems sensible to use $t = n$ in (2.9).

The test statistic r then becomes the so-called *likelihood ratio*,

$$r = \frac{L_n(\mu_0)}{L_n(\mu_1)} \tag{2.10}$$

with the likelihood function $L_n(\mu)$ of the experiment. We conclude that, when no nuisance parameters are involved, r is a somehow optimal statistic to select a hypothesis out of two.

So far, we have used the Neymna-Pearson-Lemma for the comparison of two hypotheses. As a next step, we will apply our knowledge to the construction of confidence intervals. For a given coverage, we define the optimal method to construct confidence intervals as the one that minimizes the expectation value of the intervals measure.

If we knew μ', the true value of μ, we could define for every μ the test statistic $r_\mu = \frac{L(n|\mu)}{L(n|\mu')}$ and the acceptance set A_μ as the area on the right of the cut value c_μ which depends on the desired coverage. According to the lemma of Neyman and Pearson, we thereby maximize the power for every μ with $\mu \neq \mu'$, i.e., we minimize the probability that $\mu \neq \mu'$ is in the interval[5]. Unfortunately, since we don't know μ' we have to approximate it. The usual choice is the Maximum Likelihood estimator $\hat{\mu}$. This defines a new test statistic:

$$R(n) = \frac{L_n(\mu)}{L_n(\hat{\mu})} \ . \tag{2.11}$$

$R(n)$ is called the *Maximum Likelihood Ratio*, because the denominator is now simply given by the maximized likelihood. Just like before, the acceptance set A consists of all n with $R(n) > c_\mu$. The *unified approach*, which is discussed in (3.4) is based on this test statistic.

2.9 Nuisance Parameters, Wilks' Theorem and the Profile Likelihood Ratio

Above we made the assumption that the probability distribution of the experimental outcome n is solely dependent on the parameter of interest μ. In reality, particle physicists face a huge number of additional unknown parameters. The luminosity of the particle beam is one of the most prominent examples. Despite the fact that one is usually not interested in measuring them, their influence is often crucial for the outcome of the experiment. They are therefore called *nuisance parameters* and are denoted θ. Denoting only one of them, the probability density function becomes:

$$f(n|\mu, \theta) \tag{2.12}$$

[5] It is plausible that this should in most practical cases also minimize the expectation value of the measure.

In order to make inferences about μ, it is usually necessary to conduct additional measurements that constrain the nuisance parameter. This means that n now represents both the primary measurement and the constraining measurements on the nuisance parameters.

To illustrate this, consider a counting experiment with signal μ and background θ. Events must be counted not only in the signal region but also in the background region to constrain θ. Hence n represents both event counts n_{sig} and n_{bg}. With $P(N|\nu)$ as a Poisson function with mean ν, the PDF of the experiment can be written as $f(n|\mu,\theta) = P(n_{sig}|\mu+\theta)\cdot P(n_{bg}|\theta)$.

The most natural approach is to estimate and narrow down both μ and the nuisance parameters. But this would drastically weaken the constraints we can derive for μ. Roughly said, one could reject μ only if it was outside the confidence set for all θ. In fact, we are not interested in the true value of θ. Hence it would be advisable to find a method that avoids statements about it and instead concentrates on tight confidence intervals for μ.

This is achieved by the *profile likelihood ratio*, a test statistic that is an extension of the likelihood ratio discussed above:

$$\tilde{R}_\mu(n) = \frac{L_n(\mu,\hat{\hat{\theta}})}{L(_n\hat{\mu},\hat{\theta})} \ . \tag{2.13}$$

In the denominator both μ and θ are optimized such that the likelihood reaches its maximum. In the numerator, $\hat{\hat{\theta}}$ denotes the value of θ that gives the biggest likelihood for fixed parameter μ. Possible values of $\tilde{R}$ lie in the interval $[0,1]$. For convenience, we define a transformed test statistic t_μ as:

$$t_\mu = -\ln\left(\frac{L_n(\mu,\hat{\hat{\theta}})}{L_n(\hat{\mu},\hat{\theta})}\right) = -\ln(\tilde{R}_\mu) \ . \tag{2.14}$$

In case that the index value of μ represents the true value of the parameter $\mu = \mu'$, the probability distribution of this test statistic becomes asymptotically independent from the form of the likelihood function and from the true value of the nuisance parameter θ.[6] Hence, the acceptance set for a given coverage is also independent of μ. *Wilks' theorem* states that if the PDF of the experiment satisfies certain conditions[7], this test statistic t_μ follows *in the large sample limit*[8] a χ^2*-distribution*. The number of degrees of freedom of the distribution is given by the number of parameters of interest, that is, the dimension of the vector μ.

$$g(t_\mu|\mu) = \chi^2_{dim(\mu)}(t_\mu) \tag{2.15}$$

[6] Note that we still need to know the likelihood function to be able to calculate the statistic from an experimental outcome.

[7] In practice, this also implies some constraints on θ. Assume a Gaussian distribution with mean $\mu+\theta$. In this case, any measurement would give $t_\mu = 0$.

[8] We stated above that in our notation, n can represent a vector of several independent and identically distributed random variables. For Wilks' theorem to apply, this has to be the case and the dimension of n must be sufficiently high.

3 Various Methods of Constructing Confidence Intervals

3.1 Introduction

The profile likelihood is not the only tool to create confidence intervals and depending on the purpose of the experiment it might not always be the best. In the following, we will discuss and practically demonstrate two basic alternative methods: The *Neyman construction* and the *unified approach*. Afterwards, we define specializations of the profile likelihood statistic t_μ and examine their probability distribution functions with the help of Monte-Carlo simulated data.

3.2 The Neyman Construction

This method was developed by Jerzy Neyman [Ney37]. It relies on a fixed, not further specified scalar test statistic $t(n)$ and distinguishes *central confidence intervals* from *upper* or *lower limits*. For the former, the interval $A_\mu = \left[a_{l_\mu}, a_{r_\mu}\right]$ is constructed centrally, so that

$$\frac{\gamma}{2} = \int_{-\infty}^{a_{l_\mu}} dt\ f(t(n)|\mu) = \int_{a_{r_\mu}}^{\infty} dt\ f(t(n)|\mu)\ . \tag{3.1}$$

For an upper limit, the interval is shifted to the right, then being defined by:

$$a_{l_\mu} = -\infty \text{ and } \gamma = \int_{a_{r_\mu}}^{\infty} dt\ f(t(n)|\mu)\ \ . \tag{3.2}$$

Looking at figure (3.1), one can see that these definitions give central or upper limits respectively . Acceptance Intervals for lower limits are definded analogously.

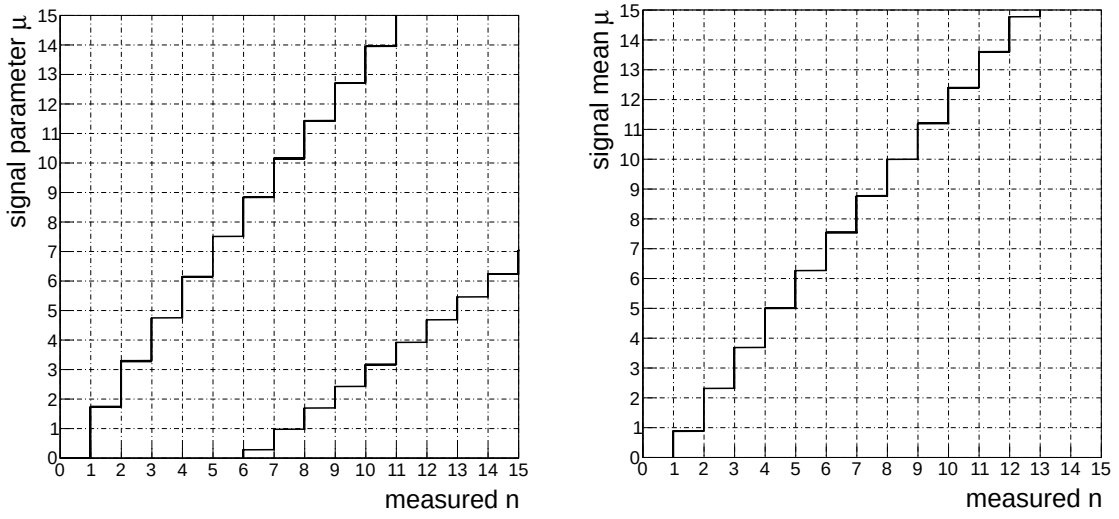
Cowan [Cow98] gives a proof for the validity of the Neyman construction that requires a_{l_μ} and a_{r_μ} to be monotonic functions of μ. Looking at the argument given in (2.6.1) we see that this is not really required. However, it is desirable because physicists prefer to state a connected Confidence Interval, which is only guaranteed if the functions are monotonic. It is then possible to define the upper and lower bound of the interval as a function of t: $\mu_l(t)$, $\mu_r(t)$

To give an example for the construction of these belts, a simple counting experiment was chosen. The expectation value ν is given by the signal s, multiplied with a strength parameter μ, plus the background expectation b: $\nu = \mu s + b$. The test statistic t is simply the measured count n, which is Poisson distributed around ν:

$$P(n|\nu) = \frac{\nu_\mu^n}{n!} e^{-\nu_\mu} \ . \tag{3.3}$$

While s and b treated as constants, μ takes the role of the parameter of interest. It is assumed that μ is always bigger than zero.

The fact that n is discrete makes it impossible to choose an interval A_μ which exactly solves (3.1) or (3.2). The central acceptance interval was thus chosen such that the cumulative probability below and above its borders is each at most equal to the desired value. This adds some conservatism, resulting in confidence intervals with a slightly higher coverage.



(a) Central Limit (b) Upper Limit

Figure 3.1: Confidence belts based on the Neyman construction at the 90% confidence level. The data was simulated for signal mean $s = 1$ and background expectation $b = 3$

Instead of directly evaluating the integral of 3.3, we chose to simulate the experiment using the built-in Monte-Carlo generator of the RooFit [VK03] toolkit. The parameters of the plots above have been chosen in a way that allows easy comparison between the resulting graphs and the plots given in the paper to the unified approach [FC98].

3.3 Flip-Flopping

Upper limit, lower limit and central confidence intervals should be chosen depending on the objective of the experiment. Lower limits are for example useful for rejecting a background-only hypothesis, whereas central confidence intervals help to find the actual strength of the signal. When conducting

an experiment, one might be tempted to decide for the preferred kind of confidence interval only *after having seen the data*. This leads to an incorrect coverage of the resulting confidence intervals, referred to as flip-flopping [FC98]: If the experiment was conducted indefinitely, the choice of a method was made depending on its outcome and the resulting confidence intervals were collected in a set $\mathcal{I}$, the fraction of elements in $\mathcal{I}$ that contain the true value of μ would in general differ from the defined confidence level. Feldman and Cousins provide a graphic explanation for this effect when motivating the unified approach in [FC98]. The incorrect coverage still holds if one considers only the subset $\mathcal{J} \subset \mathcal{I}$ of a certain kind (e.g. upper limits) of confidence intervals.[1]

One might argue that the discussion above treats a problem that does not really exist. One could say that it is enough to define an abstract coverage for a hypothetical "correct" implementation of the repeated experiment (namely without flip-flopping) and that it does not matter how the experiment was carried out in reality. However, if we want to estimate the expected fraction of publications whose stated confidence intervals really cover the true value, we need to know the coverage for the real experiment as opposed to any hypothetical implementation. It is thus desirable to truly avoid flip-flopping.

3.4 The Unified Approach

After pointing out the problem of flip-flopping, Feldman and Cousins suggest a new method of constructing confidence belts, the *unified approach*. It is aimed to give minimal confidence intervals in any situation, thus eliminating the need to flip-flop between different limit definitions. The unified approach differs from the Neyman construction only in the way the acceptance intervals A_μ are selected. It is here defined as the subset that contains all values of t with a likelihood ratio greater then c_μ:

$$A_\mu = \left\{ t : \frac{L_t(\mu)}{L_t(\hat{\mu})} > c_\mu \right\} \ , \tag{3.4}$$

with c_μ such that the desired confidence level γ is met:

$$P(A_\mu) = \gamma \ . \tag{3.5}$$

The Poisson PDF of the RooFit environment was used to directly calculate the likelihood ratio. Therefore, no Monte Carlo simulations were necessary. Feldman and Cousins restrict $\hat{\mu}$ to positive values, which is reasonable if the true value μ' of μ is known to be greater than zero. This limitation on $\hat{\mu}$ prevents the upper confidence belt from crossing the x-axis, thus avoiding the problem of a possibly empty confidence interval, something that is not guaranteed for the central interval given by the Neyman construction.

For the case of no nuisance parameters and if $t(n) := n$, the unified approach is equivalent to the $\tilde{t}_\mu$

[1] Proof: $\mathcal{I}$ is the disjunct union of subsets of different kinds of confidence intervals. If the coverage for all of them was correct, the coverage of $\mathcal{I}$ would be correct as well.

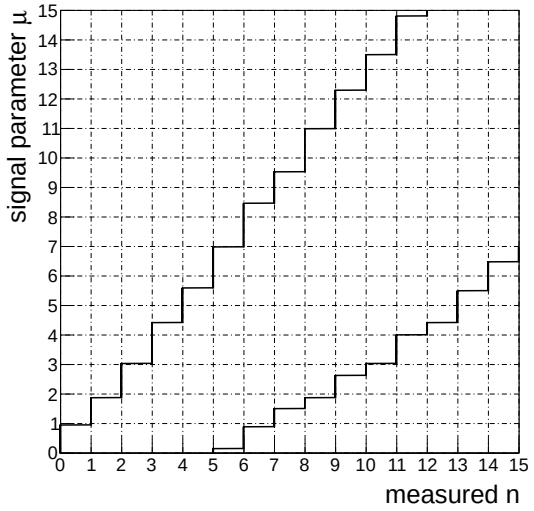


Figure 3.2: Confidence belts based on the unified approach for 90% confidence level. Once more, mean $s = 1$ and $b = 3$ were used as parameters.

statistic of the next chapter. If $\hat{\mu}$ can also be negative, the statement holds for the t_μ statistic.

However, the Unified Approach is not in all cases a good replacement for the Neyman construction, especially if one wants to set an upper limit. In the next chapter, we will discuss statistics that both make use of the profile likelihood ratio and are designed for setting lower or upper limits.

3.5 Likelihood-based Test Statistics and their Asymptotic Formulas

3.5.1 Overview

In this section, we will introduce some variations of the likelihood ratio test statistic t_μ which was introduced in (2.9). The statistics were first defined and examined in a paper [CCGV11] by Cowan, Cranmer, Gross, and Vitells. Each of them is designed to deal with special problems like the setting of upper or lower limits. They therefore combine the advantages of the Neyman construction with those of the profile likelihood ratio. For all of them, it is possible to find their probability distributions in the large sample limit by using a theorem [Wal43] of Wald that is a generalization of Wilks' theorem. The software that was written for the combination in chapter 4 relies on these asymptotic formulas. With the help of a Monte-Carlo experiment, we will simulate the real probability distributions of the test statistics and examine the agreement between them and their asymptotic predictions.

3.5.2 The test statistics

Test statistic t_μ

This is the logarithmic profile likelihood statistic that was already defined in equation (2.14). As discussed above, its motivation stems ultimately from the Neyman-Pearson lemma. Remember that the acceptance set A_μ for the likelihood ratio $R(n)$ consisted of all n such that $R(n) > c_\mu$. If we apply the same criterion for the profile likelihood ratio, μ is accepted if and only if $t_\mu(n)$ falls below a threshold t_0, which depends on the desired coverage γ:

$$\int_0^{t_0} dt\ g(t_\mu|\mu) = \gamma\ . \tag{3.6}$$

Small values of t_μ indicate good agreement between the data and the hypothesis μ. The statistic becomes zero if μ is equal to the maximum likelihood estimator $\hat{\mu}(n)$. The confidence interval is then given by all μ such that

$$\mathcal{I} = \{\mu : t_\mu(n) < t_0\} \quad . \tag{3.7}$$

If the likelihood function is symmetric around $\hat{\mu}$, the t_μ statistic gives confidence intervals that are also symmetric around $\hat{\mu}$.

Test statistic $\tilde{t}_\mu$ for $\mu' \geq 0$

It often happens that the unknown parameter is physically constrained to positive values.[2] It is possible to incorporate this knowledge into the test statistic by restricting $\hat{\mu}$ to positive values[3]:

$$\tilde{t}_\mu = \begin{cases} -2\ln \frac{L(\mu,\hat{\hat{\theta}}(\mu))}{L(\hat{\mu},\hat{\theta})} & \hat{\mu} \geq 0 \\ -2\ln \frac{L(\mu,\hat{\hat{\theta}}(\mu))}{L(0,\hat{\hat{\theta}}(0))} & \hat{\mu} < 0 \end{cases} \quad . \tag{3.8}$$

As a result, an experimental outcome that would suggest $\hat{\mu} < 0$ will not result in a rejection of small values of μ. Because of (3.6), this comes with a higher tendency to reject small μ if $\hat{\mu} > 0$. If the true value μ' is much bigger than zero, $\hat{\mu} > 0$ will always hold, and $\tilde{t}_\mu(n)$ becomes equivalent to $t_\mu(n)$.

[2] This can be generalized to any half-bounded interval.

[3] Replacing μ with 0 is motivated by the assumption that the profile likelihood function has a single local maximum at $\hat{\mu} < 0$, such that 0 is then the value that maximizes it if $\hat{\mu}$ is restricted to positive values

Setting upper limits: test statistic q_μ

When setting upper limits, an assumed μ shall only be rejected if the experimental result suggests that $\mu' < \mu$. Since μ' is unknown, $\hat{\mu}$ is used as an approximation:

$$q_\mu = \begin{cases} -2\ln\frac{L(\mu,\hat{\hat{\theta}}(\mu))}{L(\hat{\mu},\hat{\theta})} & \hat{\mu} \leq \mu \\ 0 & \hat{\mu} > \mu \end{cases} . \tag{3.9}$$

With this setting, μ will always be accepted if $\hat{\mu} > \mu$, leading to an upper limit on μ. As a consequence, the q_μ test statistic has more power in rejecting a $\mu > \hat{\mu}$. Just like for the t_μ statistic, there also exists a variation $\tilde{q}_\mu$ of q_μ that can be applied if μ' is naturally greater or equal zero.

Lower limits and the q_0 test statistic

Analogously to the q_μ statistic for upper limits, one can define a test statistic that is specialized on setting lower limits. However, Cowan, Cranmer, Gross and Vitells [CCGV11] do this only for the special case $\mu = 0$ which is important in particle physics for the rejection of a background-only hypothesis. In this special case, the lower-limit statistic is defined as:

$$q_0 = \begin{cases} -2\ln\frac{L(\mu,\hat{\hat{\theta}}(\mu))}{L(\hat{\mu},\hat{\theta})} & \hat{\mu} \geq 0 \\ 0 & \hat{\mu} < 0 \end{cases} . \tag{3.10}$$

Note that q_0 is not q_μ for $\mu = 0$. This confusing nomenclature was chosen in [CCGV11] for formal reasons.

3.5.3 Remarks

The $\tilde{q}_\mu$, $\tilde{t}_\mu$ statistcs are a special case of a more generic approach: A known limitation on μ' should be incorporated into the likelihood function by restricting its domain to physically possible values of μ'. With the help of Wald's approximation[Wal43], one can try then to find an asymptotic formula. On the other hand, if we want to use t_μ, q_μ or p_μ, we have to make sure that the likelihood function is defined for all μ (and θ).

The asymptotic formulas for the test statistics discussed here can be found in [CCGV11]. The most important distribution for every statistic is the one for the case of $\mu = \mu' : g(t_{\mu'}|\mu', \theta)$, because it is used for the construction of confidence intervals. Since μ' and θ' are unknown, it is desirable that $f(t'_\mu|\mu', \theta')$ does asymptotically not depend on these parameters. This is the case for the t_μ, q_μ and q_0 test statistics but only approximately for $\tilde{t}_\mu$ and $\tilde{q}_\mu$, because for the latter the asymptotic formulas depend on the variance of μ.

3.5.4 Monte-Carlo evaluation

With regard to the combination in chapter 4, an application was written that is able to calculate any of these test statistics out of measured data, find its P-Value[4] and the corresponding confidence intervals. For the latter two, it makes use of the asymptotic formulas mentioned above. The program accepts an arbitrary experiment ("model") and can simulate the resulting distribution of the test statistic by means of a Monte-Carlo study. A resulting plot with both the simulated and the predicted distribution can be used to investigate how well the prerequisites of Wald's approximation are met and whether the asymptotic formulas are applicable.

After testing the application with different models, the results are now demonstrated by the same counting experiment that was also used in [CCGV11]. This allows an easy comparison to the plots published in [CCGV11].

Cowan, Cranmer, Gross, and Vitells use an analytical formula to directly calculate $\hat{\mu}$, $\hat{\theta}$ and $\hat{\hat{\theta}}$ for the considered counting experiment. Hence, the calculation of the test statistics can be done without any numerical fits, enabling very fast Monte-Carlo simulations. In contrast, the generic implementation used here is able to deal with any possible experiment, but also relies on the maximum likelihood fit which heavily increases the complexity of the calculation.

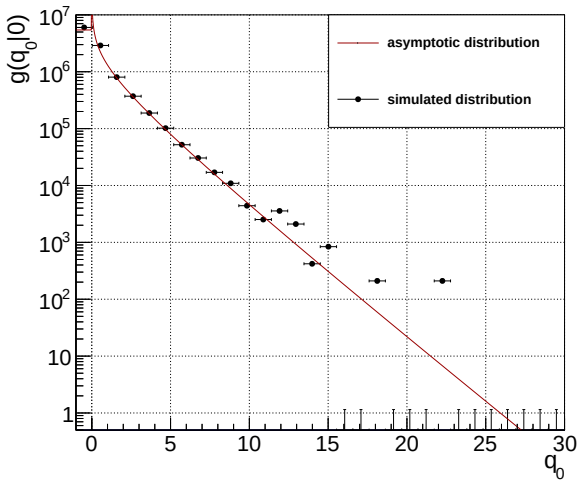
The discovery significance $Z = 5$ corresponds to the P-Value $6.004 \cdot 10^{-7}$. For one expected count in this region, about 2 million events would need to be simulated. The graphs given in the following are simulated with about 10 million events. Although all test-statistics given in [CCGV11] were implemented and tested as a part of this work, only a small selection of graphs is presented here. We concentrate on the q_0 statistic because it is the one that is primarily discussed in [CCGV11]. With regard to the combination in the next chapter, the graphs for the t_μ statistic will be presented briefly.

Figure (3.3(a)) shows the distribution of the test statistic q_0. The underflow-bin is used to express the delta-value of the PDF at $q_0 = 0$. The asymptotic PDF fits the simulated data well up to $q_0 = 11$. This can also be seen in the figure (3.3(b)) which shows the important P-value distribution. In the interval $10 < q_0 < 15$, the P-Value is underestimated, leading to incorrect confidence intervals. For higher even values, the approximation seems not to hold any more. A significance of $Z = 2$ corresponds to the P-value $P = 0.0455$. In this order of magnitude, the correspondence s still acceptable.

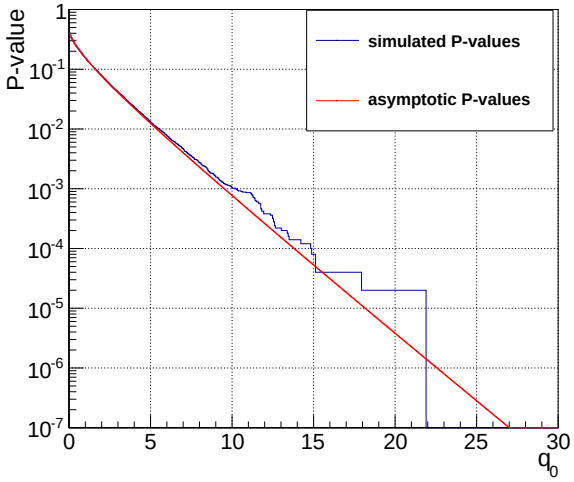
The empty bins in figure (3.3(a)) $15 < q_0 < 20$ suggest that the observed deviations are not caused by random fluctuation. They are also not caused by the discreteness of the Poisson distribution, since the same phenomenon occurs in figure (3.4(a)) that was created by replacing the Poisson functions with their Gaussian approximations.

Possible inaccurancies during the fit were researched by using the analytical formulas given in [CCGV11] to circumvent the process. Besides the fact that the obtained distribution is slightly shifted to smaller values, it still shows an empty bin at $q_0 = 16$. On the other hand, this is not seen by Cowan, Cranmer, Gross, and Vitells, who use the same method with a slightly different binning. These results

[4] In this case, the P-Value for a fixed value of the test statistic is the probability to obtain the same or a higher value of the statistic under a given hypothesis μ.

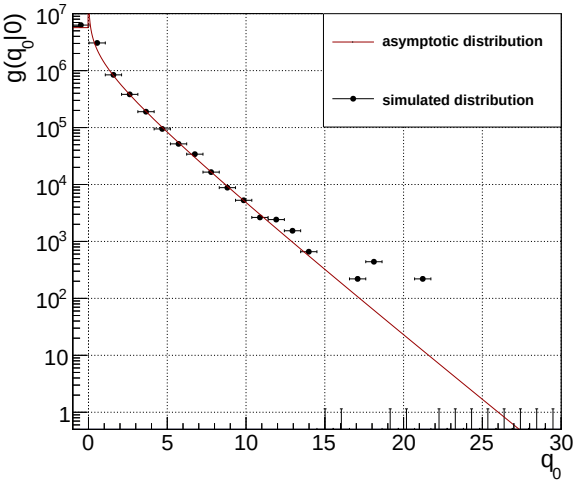


(a) Probability density funcion. The asymptotc formula is normalized to the amount of events in the histogram.

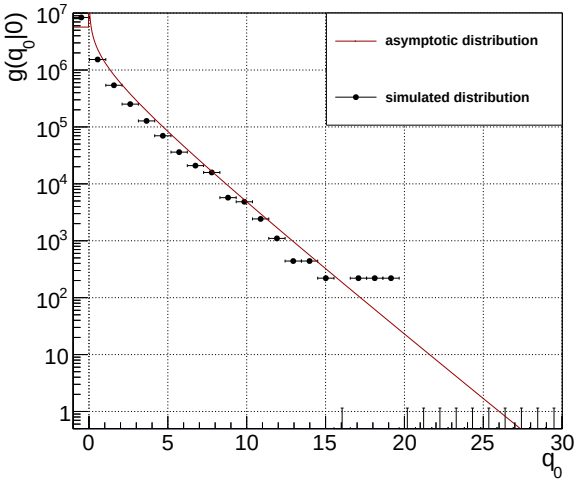


(b) P-value distribution. It has a saltus at $q_0 = 0$.

Figure 3.3: Plots of the q_0 test statistic for $\mu' = 0$ and $\theta = 10000$. $10.5 \cdot 10^6$ events were simulated.



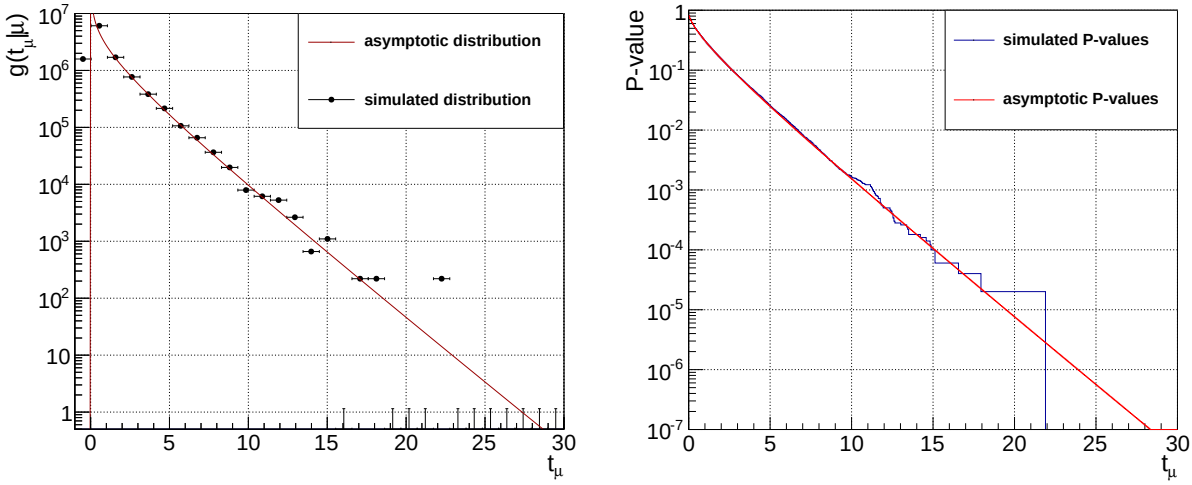
(a) PDF of q_0 for the same setting as in figure (3.3(a)), simulated with Gaussians replacing the Poisson terms.



(b) PDF of q_0 obtained by a simulation that avoided the fit process.

Figure 3.4: Reference plots for $\mu' = 0$ and $\theta = 10000$. $11 \cdot 10^6$ events were simulated.

indicate that the observed characteristics might occur in the process of simulation. It implies that the accuracy of the calculation of the test statistics might be higher than suggested by figure (3.3). The



(a) Probability density function. The asymptotc formula is normalized to the amount of events in the histogram.

(b) P-value distribution

Figure 3.5: Plots of the t_μ test statistic for $\mu = \mu' = 0$ and $\theta = 10000$. $11 \cdot 10^6$ events were simulated.

distributions simulated for the t_μ statistic can be seen in figure (3.5). While the histograms principally show the same characteristics as for the q_0 statistic, the agreement between the asymptotic and the simulated P-values is good up to $P = 10^{-3}$, corresponding to a significance higher than $Z = 3$. Hence, it is reasonable to use this test statistic for the construction of a confidence interval on the combined cross section in the next chapter.

4 Combining Measurements of the $Z \to \tau\tau$ Cross Section

4.1 Overview

The decay of the Standard Model gauge boson Z to τ leptons via the weak interaction is a rather well-understood process. Due to the importance of τ leptons in the search for new physics at the Large Hadron Collider, it constitutes an important background for many measurements. Therefore, a precise knowledge of the $Z \to \tau\tau$ cross section is a crucial prerequisite for many experiments conducted at the ATLAS detector [AAA+08].

In February 2012, an conference note [Con12] was published by the ATLAS collaboration, presenting the results of an analysis of the $Z \to \tau\tau$ cross section. The data include three different final states, resulting in three measured cross sections, one for each sub-channel. The combination of these three channels was done with the help of the BLUE [LGC88] method. As an alternative approach, we will use the method of maximum likelihood to combine the results and find a common estimator for the $Z \to \tau\tau$ cross section. Uncertainties will be estimated using the t_μ test statistic (2.14). Finally, we will compare our results to those published by the ATLAS collaboration and discuss possible reasons for the observed differences.

4.2 Sketch of the Applied Event Selection

Due to its short mean life time of about $3 \cdot 10^{-13} s$, the tau lepton decays before reaching the detector. Therefore, the final states used to detect a $Z \to \tau\tau$ event are defined by the decay modes of the τ lepton.

The τ lepton decays only via the weak interaction. It is the only lepton with sufficient energy for the creation of a quark- antiquark pair, resulting in a hadronic jet in the detector which will be referred to as τ-jet. Other decay products are electron and muon:

$$\tau \to \nu_\tau + q\bar{q}$$

$$\tau \to \nu_\tau + e + \bar{\nu}_e$$

$$\tau \to \nu_\tau + \mu + \bar{\nu}_\mu$$

The following paragraphs follow closely the discussion in [Con12].

Since the Z boson is uncharged, the two resulting τ leptons have opposite charge, leading to oppositely charged final state leptons and τ-jets. Neglecting the resulting neutrinos and omitting the distinction between particle and antiparticle, the following final states are possible: e-μ, μ-h, e-h and e-e, μ-μ, h-h. Here, h denotes a hadron. The detection of the latter three is complicated by the large Z boson background in the leptonic channels and Multijets in the hadronic one. The reconstruction of the $Z \to \tau\tau$ cross section therefore concentrates on the first three channels. In the following, they will be referred to as $\tau_e\tau_\mu$, $\tau_\mu\tau_h$ and $\tau_e\tau_h$. "Lepton" as well as "l" will refer to electrons and muons only.

The lepton is typically isolated from multijets. The τ-jet is expected to be highly collimated and to consist of an odd number of charged hadrons together with possible neutral particles. The leptonic final state includes two isolated leptons. Their transverse energy is usually lower than the ones of those produced in $Z \to ll$ events, because of the undetected neutrinos which are produced in all channels by the decaying τ and which carry away energy.

The event selection process can roughly be divided into acceptance cuts, triggering, object identification and event selection. In the analysis at hand, a muon trigger was used for the $\tau_\mu\tau_h$ and $\tau_e\tau_\mu$ decay modes, while events for the $\tau_e\tau_h$ channels were selected using a combined hadronic tau and electron trigger. Used thresholds can be found in ([Con12], sec.3).

As a next step, accepted events are analyzed for interesting patterns indicating muons, electrons, jets and missing transverse energy (E_T^{miss}). Hadronic τ candidates are detected using the Boosted Decision Tree (BDT) identification method (described in [ATL11]) which applies a variety of selection criteria. Leptons are required to be isolated[1].

The definitions of the acceptance cuts applied and other parameters for the event selection can be found in [Con12].

An event becomes a candidate to be counted in an individual channel if it contains the required objects: In the $\tau_\mu\tau_h$ and $\tau_e\tau_h$ channels, an isolated lepton was required together with an oppositely charged hadronically decaying tau candidate. In these hadronic channels, the transverse mass m_T and an angle criterion are used as test statistics to further increase the purity of the sample. The definitions of the test statistics, their MC-Simulated distribution for backgrounds and signal and the cut values are given in [Con12, sec.4.3.1 and fig.2]. The angle criterion helps suppressing $(W \to l\nu) + jets$ events, while the rejection of events with high transverse mass especially in particular reduces diboson background. After applying all these criteria in the hadronic channels, a final cut on the visible mass (the invariant mass of the lepton and the tau candidate) is applied in both channels, requiring it to be within 35 and 75 GeV. This reduces $Z \to ll$ background events.

In the $\tau_e\tau_\mu$ channel, candidates are selected by requiring exactly one electron and an oppositely charged muon. The same a angle criterion is applied, reducing the W boson background as well as the one caused by $t\bar{t}$ events. The sum of the transverse energy of all event objects (including the missing one) is required to stay below a threshold value, further suppressing the $t\bar{t}$ background. Again, the

[1] For quantitative information and formulas, see [Con12, sec. 4.2]

invariant mass is calculated and required to be within the same range as in the $\tau_\mu\tau_h$, $\tau_e\tau_h$channels. The precise definitions and cut values can be found in [Con12, sec. 4.3.2.].

4.3 Background Processes and Monte-Carlo Estimations

W bosons create background indirectly by faking tau decays via $W \to l\nu$, or directly via $W \to \tau\nu$. In the $\tau_e\tau_\mu$ channel, their influence is found to be minor and was estimated by MC- simulations. In combination with a jet faking a hadronic τ candidate, the W boson backgrounds become important in the two $\tau_l\tau_h$ final states. Since the Monte-Carlo (MC) simulations for these background show a systematic overestimate, they have to be corrected by a normalization factor which is calculated by a W boson enriched control region. This data-driven correction leads to additional systematic uncertainties.

$Z \to ll$ decays can create a background event in the $\tau_\mu\tau_e$ channel if one of the leptons is misidentified with the wrong flavour. This influence is simulated via MC. In the hadronic final states, background can be created by one lepton faking a tau candidate or by an additional jet that is misidentified as a tau candidate. These two effects are considered separately: The former is corrected in the $\tau_e\tau_h$ channel via a MC-simulation that is itself corrected with the help of a $Z \to ee$ tag-and-probe study[ATL11]. The latter is treated in both hadronic final states, again via MC and a Z -enriched control region.

Diboson and $t\bar{t}$ backgrounds are estimated by Monte-Carlo simulation.

Multijet background

Multijets are the dominant background in a proton-proton collider. A pair of real or fake leptons from a jet can fake a $\tau_\mu\tau_e$. If, instead of a second lepton, a jet is incorrectly indentified as a tau candidate, this may fake a $\tau_l\tau_h$ event. Due to high uncertainties in the simulation of these QCD events via a Monte-Carlo, the multijet background is estimated with the help of a data-driven ABCD-method[Con12]. The regions are distinguished by two criteria: Whether the measured leptons have same or opposite charge and whether the leptons fulfill the isolation requirements. This method leads to additional systematic uncertainties. For more information, please refer to [Con12, sec. 5]

Table (4.1) shows the summarized background and signal event expectations in comparison with the observed event count N for all channels after the full selection.

Table 4.1: Expected number of background and signal events in comparison with the observed event count N in each final state. The stated uncertainties are statistical only.

	$\tau_\mu\tau_h(1.55fb^{-1})$	$\tau_e\tau_h(1.34fb^{-1})$	$\tau_e\tau_\mu(1.55fb^{-1})$
Total backgroud	793 ± 34	449 ± 35	56 ± 8
$\gamma* / Z \to \tau\tau$	4544 ± 49	2029 ± 25	981 ± 26
N	5184	2600	1035

4.4 Systematic Uncertainties

The measurements in the three final states resulted in high event counts. For this reason, the influence of statistic deviations becomes small compared to the systematic uncertainties on the acceptance A_Z, on the efficiencies which are represented by the *experimental correction factor* C_Z and on the remaining background processes. A correct estimation of their influence is crucial for the quality of the combination. Table (4.2), which was taken from [Con12], summarizes the systematic uncertainties

Table 4.2: Taken from [Con12]. Relative uncertainties in percent on the total cross section measurement. A check mark in the last column indicates full correlation.

Systematic uncertainty	$\delta\sigma/\sigma$ (%) $\tau_\mu\tau_h$	$\delta\sigma/\sigma$ (%) $\tau_e\tau_h$	$\delta\sigma/\sigma$ (%) $\tau_e\tau_\mu$	Correlation
Muon efficiency	1.7	-	1.5	✓
Electron efficiency	-	5.0	6.0	✓
Muon resolution	< 0.05	-	< 0.05	✓
Electron resolution	-	0.1	0.2	✓
Jet resolution	-	-	1.7	–
τ ID efficiency	5.2	5.2	-	✓
$e \to \tau$ misidentification rate	-	0.2	-	–
Energy scale	8.2	9.3	4.5	✓
τ trigger efficiency	-	4.7	-	–
W normalization factor	<0.05	<0.05	-	–
Z normalization factor	<0.05	<0.05	-	–
Multijet estimation	0.8	1.3	0.4	✓
Background theor. cross section	0.1	0.2	0.2	✓
Monte Carlo statistics	1.2	1.4	2.9	–
A_Z uncertainties	3.1	3.4	3.2	✓
Total systematic unc.	10.4	13.2	8.9	
Luminosity uncertainty	3.7	3.7	3.7	✓
Statistical uncertainty	1.6	2.4	3.3	–

applied by the authors. They will also be used for the ML-combination in the next chapter.

The muon and electron efficiency systematics include the uncertainty on the chain process of triggering, reconstruction and identification cuts for the leptons. Data-driven methods were used to estimate these efficiencies and to correct related Monte-Carlo simulations. The *lepton resolution* uncertainties refer to the limited detector resolution and were found to have very small influence. The *jet resolution* influences the $\tau_e\tau_\mu$ channel because a direct cut on E_T^{miss} is applied here. The τ *identification efficiency* of the BDT method was estimated with the help of data. The same holds for the $e \to \tau$ *misidentification rate* which was corrected for the Monte-Carlo simulation. *Energy scale* systematics origin from uncertainties on the calibration of the calorimetric detector devices. They are treated as fully correlated among the different final states. The τ *trigger efficiency* is measured in data depending on p_T and applied on the Monte-Carlo as a weighting factor. Its influence is limited to the $\tau_e\tau_h$ final state, because the other channels do not rely a the *tau* trigger. The *W and Z normalization factors* in the $\tau_l\tau_h$ channels are mainly affected by statistical uncertainties in the control regions used for their calculation. The *multijet estimation* is influenced by statistical and systematic uncertainties

in the control regions of the ABCD method. *Background theoretical cross section* sums up uncertainties on the input variables of the Monte-Carlo simulations. They are correlated and their influence on the final cross section is estimated. The bullet point *Monte Caro statistics* refers to all statistical uncertainties caused by simulations of A_Z, C_Z and the background events. Since they are computed seperately for each final state, no correlation is expected. A_Z *uncertainties* include all systematic uncertainties in conjunction with the applied acceptance cuts. They are caused by limited knowledge of the Monte-Carlo simulated signal process. The *luminosity* has a correlated and equal influence on all channels.

4.5 Cross Section Calculation

The event count in the i-th of the three channels will be denoted as $\hat{N}_i$.[2][3] It is assumed that both $\hat{N}_i$ and the amount of background events are Poisson distributed around their mean values N_i and M_i. The true value Σ[4] for the $Z \to \tau\tau$ cross section is then given by

$$\Sigma = \frac{N_i - M_i}{B_i \cdot A_{Zi} \cdot C_{Zi} \cdot \mathcal{L}} \ . \tag{4.1}$$

i channel-index $i = 1, 3, 3$

B_i branching fraction of the i-th channel

$\mathcal{L}$ luminosity

The measured values for the different final states are stated in [Con12] as:

Final State	Total cross section $\tilde{\Sigma}_i$
$\tau_\mu\tau_h$	$0.91 \pm 0.01(stat) \pm 0.09(sys) \pm 0.03(lumi)\ nb$
$\tau_e\tau_h$	$1.00 \pm 0.02(stat) \pm 0.13(syst) \pm 0.6(lumi)\ nb$
$\tau_e\tau_\mu$	$0.96 \pm 0.03(stat) \pm (0.09)(syst) \pm 0.04(lumi)\ nb$

4.6 Constructing the Likelihood Function

In order to find a combined ML - estimator for the $Z \to \tau\tau$ cross section and to calculate confidence intervals, we will now have to specify the functional form of the likelihood function. Because the systematic uncertainties cause the measured cross section to be correlated, we cannot simply apply the method given in [Cow98, chap. 6.12]. While we will have to make some assumptions and approximations, the following arguments are purely frequentist.

The final likelihood function must include the measurements of the nuisance parameters $\{\theta_k, \theta_l, ..\}$.

[2]This notation is motivated by the fact that $\hat{N}_i$ is also the ML-estimator for the true but unknown N_i.
[3]In the following, i will always indicate the instance of a variable that corresponds to the i-th channel.
[4]The usual symbol for a cross section is σ. Σ is chosen to distinguish it in this context from the standard deviation σ.

Therefore, the PDF of the "whole" experiment is a product of the Poisson functions p_i for the counting experiments in the different channels, multiplied with the PDFs h_j for the variables $\tilde{\theta}_j$ which represent direct or indirect measurements of the nuisance parameters θ_j:

$$\begin{aligned} L =& p_1(\hat{N}_1|N_1(\Sigma,\theta_1,\theta_2,...)) \cdot p_2(\hat{N}_2|N_2(\Sigma,\theta_1,\theta_2,...)) \cdot p_3(\hat{N}_3|N_3(\Sigma,\theta_1,\theta_2,..)) \cdot ... \\ & \cdot h_k(\tilde{\theta_k}|\theta_k,\theta_l,...,\tilde{\theta_l},\tilde{\theta_3},...) \cdot ... \cdot h_l(\tilde{\theta_l}|\theta_k,\theta_l,...,\tilde{\theta_3},...) \cdot ... \ . \end{aligned} \tag{4.2}$$

$\hat{N}_i$ measured count in the i-th channel

$\theta_k, ...$ nuisance parameter

$\tilde{\theta}_k$ measured value of an experiment that aims to constrain θ_k. Not necessarily a direct measurement of θ_i

In the following, we will have to transform this equation into a form that contains only known quantities. The Poisson distribution of $\hat{N}_i$ can be approximated by a Gaussian[5] with standard deviation $\sigma_i = \sqrt{N_i}$. We will also assume $\sigma_i \approx \sqrt{\hat{N}_i}$ to be known and fixed. Furthermore, all $\tilde{\theta}_k$ are assumed to be Gaussian distributed with known and fixed variance, a simplification that might sometimes be incorrect. In general, the measured values $\tilde{\theta_k}$ have a function $m(\theta_k;\tilde{\theta}_l,...)$ as their mean. This general approach allows for conditional probabilities on the nuisance parameters which frequently occur in an experimental setting. For example, this is the case for Monte-Carlo estimators $\tilde{\theta_k}$, where $(\tilde{\theta}_l,...)$ are the input parameters of the simulation and θ_k is a parameter that embodies the ignorance about the form of the function $m_{\theta_k}(\tilde{\theta}_l,...)$ that is approximated by the Monte-Carlo simulation. Other values $\tilde{\theta}_l$, like for example the estimator of the luminosity, are direct measurements of a nuisance parameter θ_l which then takes the role of the central value[6]. Defining the scale factor

$$c = B_i \cdot A_{Zi} \cdot C_{Zi} \cdot \mathcal{L} \tag{4.3}$$

and using equation (4.1), this leads to:

$$\begin{aligned} L =& G(\hat{N}_1, \Sigma \cdot c_1(\theta_k,\theta_l,...) + M_1(\theta_k,\theta_l,...),\sigma_1) \cdot G(\hat{N}_2, \Sigma \cdot c_2(\theta_k,\theta_l,...) + M_2(\theta_k,\theta_l,...),\sigma_2) \cdot ... \\ & \cdot G(\tilde{\theta_k}, m(\theta_k;\tilde{\theta}_l,...),\sigma_{\theta_k}) ... \cdot G(\tilde{\theta}_l,\theta_l,\sigma_{\theta_l}) \cdot ... \end{aligned} \tag{4.4}$$

which equals:

$$\begin{aligned} L =& G\left(\frac{\hat{N}_1 - M_1(...)}{c_1(...)}, \Sigma, \frac{\sigma_1}{c_1(...)}\right) \frac{1}{c_1(...)} \cdot G\left(\frac{\hat{N}_2 - M_2(...)}{c_2(...)}, \Sigma, \frac{\sigma_2}{c_2(...)}\right) \frac{1}{c_2(...)} \cdot ... \\ & \cdot G\left(\frac{\tilde{\theta_k} - m(\tilde{\theta}_l,...)}{\sigma_{\theta_k}}, 0, 1\right) \frac{1}{\sigma_{\theta_k}} \cdot G\left(\frac{\tilde{\theta}_l - \theta_l}{\sigma_{\theta_l}}, 0, 1\right) \frac{1}{\sigma_{\theta_l}} \cdot ... \end{aligned} \tag{4.5}$$

The factors h_i that contain only measurements of nuisance parameters are often referred to as *constraining terms*. For convenience, we will abbreviate their first arguments with new variables, φ_i.

[5] We denote a Gaussian function with argument x, mean $\bar{x}$ and standard deviation σ as $G(x,\bar{x},\sigma)$

[6] In the following calculations, $\tilde{\theta}_l$ is used as an example for a nuisance parameter of the latter kind, while $\tilde{\theta}_k$ will represent a nuisance parameter with a conditional mean.

Hence, for example:

$$\varphi_l := \frac{\tilde{\theta_l} - \theta_l}{\sigma_{\theta_l}} \qquad \text{or} \qquad \varphi_k := \frac{\tilde{\theta_k} - m(\theta_k; \tilde{\theta}_l, ...)}{\sigma_{\theta_k}} \ . \tag{4.6}$$

Treating the θ_i as constants, any statistic that does not depend on $\hat{N}_i$ can be written as a function of the new variables φ_i:

$$f(\tilde{\theta}_k, \tilde{\theta}_l, ...) = f(\varphi_k \sigma_{\theta_k} + m(\theta_k; \varphi_l \sigma_{\theta_l} + \theta_l), \varphi_l \sigma_{\theta_l} + \theta_l, ...) = \bar{f}(\varphi_k, \varphi_l, ...) \ . \tag{4.7}$$

Applying the Taylor expansion

$$\bar{f}(\varphi_k, \varphi_l, ...) = \bar{f}(0, 0, ...) + \left.\frac{d\bar{f}}{d\varphi_k}\right|_0 \varphi_k + \left.\frac{d\bar{f}}{d\varphi_l}\right|_0 \varphi_l + ... \tag{4.8}$$

and using

$$\bar{f}(0, 0, ...) = f(m(\theta_k; \theta_l, ...), \theta_l, ...) \ , \tag{4.9}$$

we get

$$\begin{aligned} f(m(\theta_k, \theta_l, ...), \theta_l, ...) = \ & f(\tilde{\theta}_k, \tilde{\theta}_l, ...) - \left.\frac{d\bar{f}}{d\varphi_k}\right|_0 \varphi_k - \left.\frac{d\bar{f}}{d\varphi_l}\right|_0 \varphi_l - ... \\ = \ & f(\tilde{\theta}_k, \tilde{\theta}_l, ...) \left[1 - \alpha_1 \varphi_k - \alpha_2 \varphi_l - ...\right] \end{aligned} \tag{4.10}$$

$$\text{with} \quad \alpha_i := \frac{1}{f(\tilde{\theta}_k, \tilde{\theta}_l, ...)} \left.\frac{d\bar{f}}{d\varphi_i}\right|_0 \ . \tag{4.11}$$

Applying all these definitions to those arguments of (4.5) that contain nuisance parameters, we are now able to give (4.4) a new form:

$$\begin{aligned} L = & G\left(\frac{\hat{N}_1 - M_1(\tilde{\theta_k}, \tilde{\theta}_l, ...)}{c_1(\tilde{\theta_k}, \tilde{\theta}_l, ...)} \left[1 - \alpha_1 \varphi_k - ...\right], \Sigma, \frac{\sigma_1}{c_1(...)} \left[1 - \beta_1 \varphi_k - ...\right]\right) \frac{1}{c_1(...)} \left[1 - \beta_1 \varphi_k - ...\right] \\ & \cdot G\left(\frac{\hat{N}_2 - M_2(...)}{c_2(...)} \left[1 - \gamma_1 \varphi_k - ...\right], \Sigma, \frac{\sigma_2}{c_2(...)} \left[1 - \delta_1 \varphi_k - ...\right]\right) \frac{1}{c_2(...)} \left[1 - \delta_1 \varphi_k - ...\right] \cdot ... \\ & \cdot G\left(\varphi_k, 0, 1\right) \frac{1}{\sigma_{\theta_k}} \cdot G(\varphi_l, 0, 1) \frac{1}{\sigma_{\theta_l}} \cdot ... \quad . \end{aligned} \tag{4.12}$$

The terms $\tilde{\Sigma_i}(\tilde{\theta}_k, ...) := \left(\hat{N}_i - M_i(\tilde{\theta}_k, ...)\right) / c(\tilde{\theta}_k, ...)$ are the measured cross sections. To clarify how the systematic uncertainties in table (4.2) should be interpreted in this context, it is important to understand how they are determined. In general, the relative uncertainty of a nuisance parameter is estimated by varying the concerning variable by one standard deviation and measuring the relative change of the statistic of interest. If necessary, authors of [Con12] symmetrized the obtained uncertainties. When examining the influence of a variable like θ_l, both the direct influence of θ_l and its indirect influence via the mean of θ_k is taken into account. Comparing with (4.7), we find that the

correct expression for the relative errors is given by:

$$\frac{f(\varphi_k,...,\varphi_i \pm 1,...) - f(\varphi_k,...,\varphi_i,...)}{f(\varphi_k,...,\varphi_i,...)} \approx \frac{\frac{df}{d\varphi_i}\Big|_{\tilde{\varphi}_1,\tilde{\varphi}_2,...}}{f(\varphi_k,...,\tilde{\varphi}_i,...)} \approx \frac{\frac{df}{d\varphi_i}\Big|_0}{f(\varphi_k,...,\varphi_i,...)} = (4.11) = \alpha_i \ . \quad (4.13)$$

The values in table (4.2) are estimated with respect to $f_i = \tilde{\Sigma}_i(\tilde{\theta}_k,...)$, but nothing is said about the systematic uncertainty on $1/c(\tilde{\theta}_k,...)$ alone. The easiest way of dealing with this is by disregarding these systematics. This leads to the likelihood function used in the fit:

$$\begin{aligned} L \approx & G\left(\tilde{\Sigma}_1\left[1-\alpha_1\varphi_k-\alpha_2\varphi_l-...\right],\Sigma,\frac{\sigma_1}{c_1(...)}\right)\cdot G\left(\tilde{\Sigma}_2\left[1-\gamma_1\varphi_k-\gamma_2\varphi_l-...\right],\Sigma,\frac{\sigma_2}{c_2(...)}\right) \\ & \cdot G\left(\tilde{\Sigma}_3\left[1-\varepsilon_1\varphi_k-\varepsilon_2\varphi_l-...\right],\Sigma,\frac{\sigma_3}{c_3(...)}\right)\cdot G\left(\varphi_k,0,1\right)\cdot G(\varphi_l,0,1)\cdot ... \quad . \end{aligned} \quad (4.14)$$

(Constant factors are omitted.)

For the systematics marked as correlated, one variable φ_i is used with different factors for each channel, resulting in one constraining Gaussian. For the uncorrelated ones, it is assumed that they have also been measured independently, hence the combination uses a separate variable φ_i and constraining Gaussian for each channel.

Equation (4.14) does not contain any explicit dependence on the original nuisance parameters θ_i any more. The influence of the nuisance parameters is incorporated in the corresponding $\varphi_i(\theta_i)$. If $\varphi_i(\theta_i)$ is a bijective function in the interesting range of parameters (this is assumed to hold), we can use (4.14) to find both the ML- Estimator for Σ and the likelihood ratio for fixed Σ. This is because its maximum value, when maximizing for (free or fixed) Σ and variable φ_i is the same[7] as the maximum of 4.4 when maximizing for the corresponding (free or fixed) N_i and variable θ_i.

The implementation of this fit is discussed in chapter 5.

4.7 Results and Discussion

The combined result together with 1σ confidence intervals, evaluated with the method of maximum likelihood, is:

$$\Sigma = 0.88\ nb \quad , \quad [0.80\ nb, 0.96\ nb]\ \ (total) \quad , \quad [0.87\ nb, 0.89\ nb]\ \ (stat)\ .$$

The cited confidence interval was estimated by finding the P-value distribution of the t_μ statistic with the help of its approximate distribution. Statistical uncertainties were estimated in the same way after fixing the nuisance parameters to their fit values $\hat{\theta}_i$. The resulting graphs are shown in figure (4.1). The visible artefacts are caused by numerical deviations during the fit but do not have any influence on the final result.

[7] apart from the omitted constant factor and errors caused by approximations

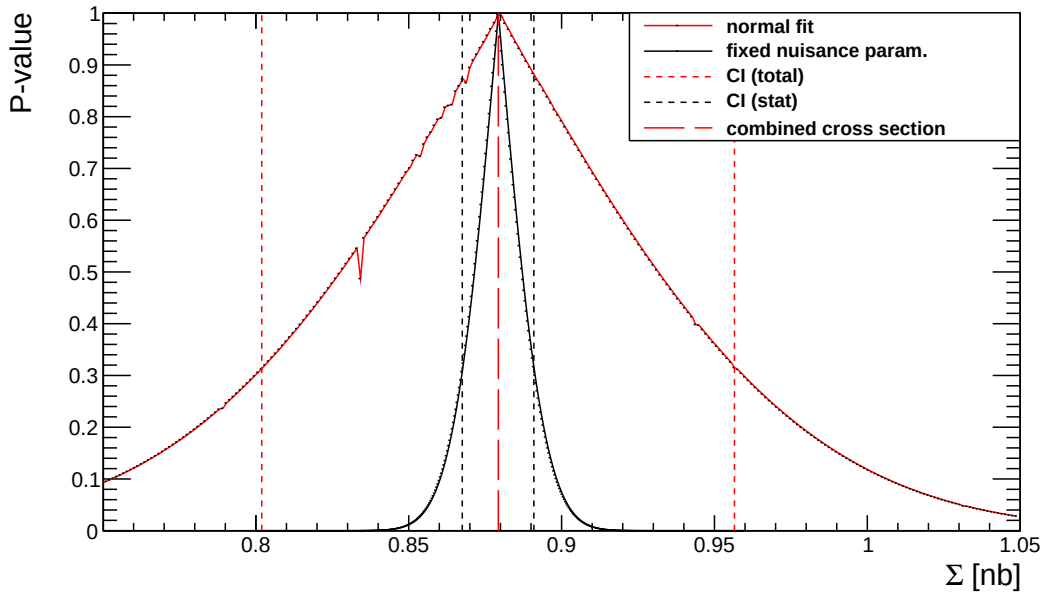


Figure 4.1: P-value distributions for free and fixed nuisance parameters. The vertical lines mark the confidence intervals and the combined cross section.

It is striking that the combined cross section is significantly smaller than each of the cross sections in the individual channels. In principle, this is not implausible since the combined result is systematically corrected, suggesting that the individual measurements are influenced by a common systematic that causes an upward fluctuation. However, the approximations described in the previous section might have influenced the result. As a prominent example, omitting the scaling factor on $\frac{\sigma_{\tilde{\theta}_i}}{c_i}$ lowers the result of the combination. This is because it allows to scale down the numerator in the exponent of the Gaussian without applying the same correction factor to the variance in the denominator. Unfortunately, the authors of [Con12] do not state relative uncertainties on c_i, but it is possible to roughly select systematics in table (4.2) with dominant influence on $\frac{1}{c_i}$[8]. After a small approximation that is necessary to ensure correct normalization, the improved likelihood function is fitted, leading to: $\Sigma' = 0.90\ nb \quad [0.83\ nb, 0.99\ nb]\,(total) \quad [0.89\ nb, 0.91\ nb]\,(stat)$.

The corresponding P-value distribution can be seen in figure (4.2). It is asymmetrical with a tendency to higher values of Σ. While this result should not be taken too seriously, it gives an impression of the influence of this effect and should be seen as a motivation to further improve the combination by using less approximations and more low-level information.

The theoretical prediction for the $Z \to \tau\tau$ cross section is $0.96 \pm 0.05\ nb$, given an invariant mass between 66 and 116 GeV. Its central value is contained in the 1σ confidence interval of the combination here. The authors of [Con12] made use of the BLUE[LGC88][Val03] method to combine the measured cross sections from the individual final states. The published combined cross section from that combination is $\Sigma = 0.92 \pm 0.02(stat) \pm 0.08(syst) \pm 0.03(lumi)\ nb$. But this result was obtained only when the authors omitted dominant systematics from the BLUE combination. The BLUE combination including all uncertainities seemed to lead to a smaller result as well.

[8] The luminosity is an example

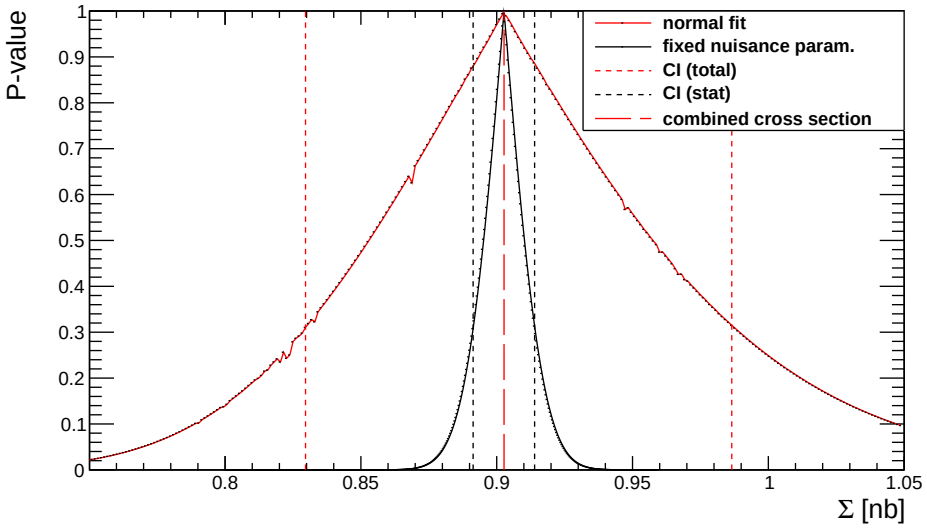


Figure 4.2: P-value distributions when approximately considering the systematic error on c_i.

The BLUE method uses Gaussian error propagation to calculate from the variances of the nuisance parameters the covariance matrix of the three statistics $\tilde{\Sigma}_i$ $(i = 1, 2, 3)$. It is thus not surprising that the result differs from the one that was obtained in the ML- combination.

5 Computation and Program Design

The calculations, fits and simulations in this work were programmed in C++. The ROOT[BR97] framework was used, with a special focus on the RooFit[VK03] toolkit. The plots for the Neyman method and the unified approach were created with rather simple, script-like programs. The simulation of the likelihood-based test statistics in chapter 3 and the combination of measurements in chapter 4 were done with the same program. The architecture bases on two generic classes that respectively represent a random experiment and a test statistic together with its asymptotic formula. To create a test statistic, one has to implement the abstract interface. Analogically, one can define an experiment by implementing the probability density function and stating the set of parameters of interest. For the combination itself, a class was written that allows it to create the PDF by defining the list of nuisance parameters together with their systematic errors. From the RooFit toolkit, the workspace class was used as an essential device to create generic PDFs and their argument parameters. It was also used to set up the probability density function of the experiment. Unfortunately, the workspace class heavily relies on string arguments and implicit operations which make it difficult to write generic code.

The RooFit toolkit automatically normalizes any given PDF with regard to a specified set of random variables. This is problematic for the combination because the Gaussians in equation (4.14) should be normalized with respect to the whole first argument, not only with respect to $\tilde{\Sigma}$. It makes it also difficult to consider errors on the normalization that are caused by the uncertainty on $\frac{1}{c_i}$. Special care had to be taken to ensure correct normalization in both cases. More generally, the RooFit classes are good for the depiction of real experiments like the double Poisson function in section 3.5.2 or the unchanged likelihood function (4.2) from the combination in chapter 4. They are less useful when the originally measured values are already processed into a more high-level result like the measured cross sections $\tilde{\Sigma}_i$. This should be seen as a motivation to apply the maximum likelihood method earlier in the evaluation of experiments.

6 Outlook

As we have seen in the evaluation of the cross section combination, the final result changes depending on the approximations that are made. Hence, they should be avoided. Unfortunately, the process of combination in this work was largely determined by the available information in [Con12]: All uncertainties were stated by their influence on $\tilde{\Sigma}$. One could facilitate and improve the combination by separately evaluating the systematic uncertainties on the background expectation M_i and the scaling factor c_i, including possible correlations. For the combination, a form of the likelihood function similar to (4.4) could then be applied. This "natural" version of the likelihood function would also avoid any normalization-related difficulties with the RooFit toolkit.

It has been stated that the Maximum Likelihood Ratio is the best statistic for constructing confidence intervals. The contraposition is that any other applied cut during the event selection weakens the theoretically possible constraints. The reason why cuts are necessary today is the limited knowledge of the likelihood function. If for any observed detector event, we were able to state the likelihood as a function of a limited set of nuisance parameters and parameters of interest, no event selection would be necessary. We could then directly find the (profile) likelihood for the total event data of the ATLAS detector to be obtained if the higgs boson existed - or not. Today this is an unrealistic suggestion. But it offers a perspective how the difficult process of event selection might change in the coming years.

7 Bibliography

[AAA+08] G. Aad, E. Abat, J. Abdallah, A. Abdelalim, A. Abdesselam, O. Abdinov, B. Abi, M. Abolins, H. Abramowicz, E. Acerbi et al., *The ATLAS experiment at the CERN large hadron collider*, Journal of Instrumentation **3**, S08003 (2008).

[ATL11] Performance of the Reconstruction and Identification of Hadronic Tau Decays with ATLAS, ATLAS-CONF-2011-152, Nov 2011.

[BR97] R. Brun and F. Rademakers, *ROOT — An object oriented data analysis framework*, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and AssociatedEquipment **389**(1–2), 81 – 86 (1997), New Computing Techniques in Physics Research V.

[CCGV11] G. Cowan, K. Cranmer, E. Gross and O. Vitells, *Asymptotic formulae for likelihood-based tests of new physics*, The European Physical Journal C - Particles and Fields **71**, 1–19 (2011), arXiv:physics.data-an/1007.1727, 10.1140/epjc/s10052-011-1554-0.

[Con12] Z -> tau tau cross section measurement in proton-proton collisions at 7 TeV with the ATLAS experiment, Technical Report ATLAS-CONF-2012-006, CERN, Geneva, Feb 2012.

[Cow98] G. Cowan, *Statistical data analysis*, Oxford University Press, USA, 1998.

[Cow07] G. Cowan, *Data analysis: Frequently Bayesian.*, Physics Today **60**(4), 82 – 83 (2007).

[FC98] G. J. Feldman and R. D. Cousins, *Unified approach to the classical statistical analysis of small signals*, Phys. Rev. D **57**, 3873–3889 (Apr 1998).

[LGC88] L. Lyons, D. Gibaut and P. Clifford, *How to combine correlated estimates of a single physical quantity*, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **270**(1), 110 – 117 (1988).

[Ney37] J. Neyman, *Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability*, Royal Society of London Philosophical Transactions Series A **236**, 333–380 (August 1937).

[Val03] A. Valassi, *Combining correlated measurements of several different physical quantities*, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **500**(1–3), 391 – 405 (2003), NIMA Vol 500.

[VK03] W. Verkerke and D. Kirkby, *The RooFit toolkit for data modeling*, ArXiv Physics e-prints (June 2003), arXiv:physics/0306116.

[Wal43] A. Wald, *Tests of statistical hypotheses concerning several parameters when the number of observations is large*, American Mathematical Society (1943).

Erklärung

Hiermit erkläre ich, dass ich diese Arbeit im Rahmen der Betreuung am Institut für Kern- und Teilchenphysik ohne unzulässige Hilfe Dritter verfasst und alle Quellen als solche gekennzeichnet habe.

Konstantin Schubert

Dresden, Mai 2012